Object-based Image Retrieval with Attention Analysis and Spatial Re-ranking

Ke Gao[1], Shouxun Lin[2], Yongdong Zhang[2] and Sheng Tang[2]

[1]Key Laboratory of Intelligent Information Processing,

Institute of Computing Technology, Chinese Academy of Sciences

Graduate University of the Chinese Academy of Sciences

Beijing, 100080,China

kegao@ict.ac.cn

[2]Key Laboratory of Intelligent Information Processing,

Institute of Computing Technology, Chinese Academy of Sciences

Beijing, 100080

China

{sxlin, zhyd, ts}@ict.ac.cn

Abstract: In this paper, a new method is proposed for object-based image retrieval. The user supplies a query object by selecting a region from a query image, and the system returns a ranked list of images that contain the same object, retrieved from a large image database. The main outcomes of this research are as follows: (1) An novel object-based image retrieval framework that integrates effective pre-treatment and re-ranking is presented, (2) a new feature filtration method based on attention analysis is proposed for pre-treatment, (3) to further improve object retrieval precision, we add an efficient spatial configuration model to re-rank the primary retrieval result using Bag of Word method. Experimental results demonstrate the effectiveness of our method.

Keywords: Object-based image retrieval, attention analysis, spatial re-ranking

1 Introduction

OBIR (Object-based Image Retrieval) is an important branch of content-based image retrieval (J.Sivic and A.Zisserman, 2003; J.Phibin et al.2007). The goal of OBIR is to find images containing desired object by providing the system a selected region of a query image. It remains a challenging problem because an

object's visual appearance may be quite different due to viewpoint, illumination, affine transformation, and even partially occlusion.

The innermost core of OBIR is how to detect and measure the similarities of object regions. Recent work in this field can be divided into two categories: one is based on image segmentation, such as Blobworld and SIMPLcity (C. Carson *et al.*1999; Wang *et al.*2001); the other is so-called BoW (Bag of Words) method, which simulates simple text-retrieval system using the analogy of "visual words" (J.Sivic *et al.*2003). BoW doesn't rely on the precision of image segmentation, and can deal with a variety of affine transformations. In consequence, it has become increasingly attractive (Qing-Fang Zheng *et al.*2006; S. Lazebnik *et al.*2006).

In Bag of Words method, affine covariant local patches (Mikolajczyk and Schmid, 2002; Matas *et al*, 2002) are detected in images, and an affine invariant descriptor (Lowe, 2004) is computed for each patch. To effectively index these high-dimensional descriptors, they are clustered into a visual vocabulary, and each patch is mapped to its closest visual word. Then an image is represented as a bag of visual words and their frequency of occurrence. Usually, they are organized as an inverted file to facilitate efficient retrieval. The benefits of this approach are as follows: first, the use of local affine covariant patches and affine invariant descriptors can effectively present an object under various image transformations, such as different viewpoint, illumination, affine transformation, and even partially occlusion; second, feature matching has been pre-computed using vector quantization, so that any particular object can be retrieved at run-time.

Although BoW method is an effective analogy between "visual words" and text words, Object-based image retrieval using "visual words" and text words-based web pages retrieval are somewhat different. The "visual words" are calculated by unsupervised clustering, and can't be understood by the user. Accordingly, "visual words" are "noisier" than text keywords in two aspects: on one hand, patch detector often returns a large number of patches which have a low signal-to-noise ratio, because only a few of them are distinguishable, so informative patches need to be picked out through a sea of background patches; on the other hand, in BoW method, the spatial information about the image-location of the visual words is ignored, which is similar to retrieve documents only by orderless letters. This will result in false matching such as "abc=cba", and reduce the retrieval precision.

To utilize the spatial relation between patches, Sivic et al. use a search area containing the 15 nearest neighbors of each matched patch, and the neighboring patch which also matches within this area casts a vote for that image. Philbin adds affine matrix verification to nearby patches using LO-RANSAC. Zheng et al. propose a visual phrase-based approach using adjacent patch pair, which is hard to satisfy in images with sparse patches, and doesn't contain the information of distance and orientation. Furthermore, the above methods don't consider the spatial neighborhood' affine transformation under different viewpoints, and rely heavily on the clustering precision of visual words.

To solve the above problems, the benefits of this paper are as follows. (1)A novel object-based image retrieval framework that integrates effective pre-treatment and re-ranking is presented, (2) a new feature filtration method based on attention analysis is proposed for pre-treatment, (3) to further improve object retrieval precision, we add an efficient spatial configuration model to re-rank the primary retrieval result using Bag of Word method.

The remainder of the paper is organized as follows. Section 2 gives a overview of our system. Section 3 describes attention analysis based feature filtration in detail. Spatial re-ranking is discussed in Section 4, and Section 5 concludes this paper.

2 Overview of the object-based image retrieval system

As the system flow chart showed in figure 1, after local affine covariant patches are detected, attention-based filtration select informative patches from them. Then Bag of words model is used to obtain the candidate image set and greatly reduce the number of images that need to be considered. It is can be efficiently implemented as an inverted file data-structure. Finally, we use spatial configuration model to re-rank the candidate image set, and improve the primary retrieval precision. Following (K.Mikolajczyk et al, 2005), we use MS (Matas, 2002) algorithm as region detector and SIFT (Lowe, 2004) to describe the regions.

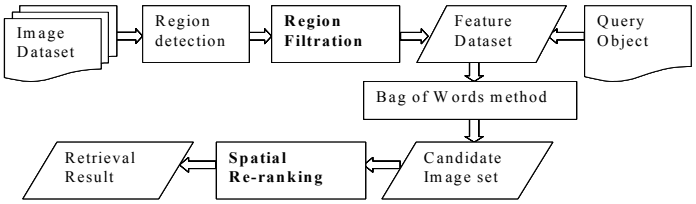

Figure 1. The flow chart of our OBIR system.

3 Attention analysis based region filtration

Although MS region detection algorithm can deal with various image affine transformation, it often generates a large number of patches which only a few of them are distinguishable, so informative patches need to be picked out through a sea of background patches. “Background patches” we mentioned here includes two kinds of patches: one kind comes from background rather than the salient region of the image; the other has little distinctive information thus can be found in both foreground object and background. Consequently, this section focuses on the

above problem and proposes a novel method to filter these affine covariant regions based on attention model and local entropy.

Our contribution lies in proposing a novel method which is well-suited to filter MS patches based on attention model and local entropy. Using attention analysis and local entropy, all patches detected in an image are ranked with its saliency, and only the most distinctive patches will be reserved. In this way, the local patch information and global image distribution are both taken into account, and the background patches can be removed effectively.

3.1 Attention model and saliency region

Attention is at the nexus between cognition and perception. While interpreting a complex scene, a human being selects a subset of the available sensory information before further processing. This region is so-called "focus of attention" (Tsotsos *et al*, 1995). (Itti *et al*, 2001) proposed a saliency-based attention model for scene analysis. In his work, a "saliency map" is generated in a bottom-up manner as a combination of these feature maps. Recently, fuzzy growing (Ma *et al*, 2003) is proposed to find all of the saliency regions for original image. Considering the calculation complexity, the number of saliency regions per image is limited to 3. Figure 2 gives an example in practical application.

(a) (b) (c)

Figure 2. Samples of saliency regions detection. (a) original images, col. (b) attention model based saliency map, col. (c) saliency regions (as figured out by yellow rectangles), col.

Given a patch X lies in any saliency regions and its grey level distribution $D = \{ d_1 \ldots\ldots d_r \}$, local entropy is defined as:

$$H_X = -\sum_{i=1}^{r} p(d_i) * \log_2 p(d_i) \tag{1}$$

Where $p(d_i)$ is the probability of pixel taking the value d_i in patch X. Informative patches often have large entropy, so we remove those patches whose entropy is less than threshold $Entropy_{low}$.

As to the patches from trees or grass which have complex texture, because they have similar intensity distribution over large ranges of scale, we use self-similarity to remove them. For simplicity the sum of absolute difference of several grey-level histograms is defined as self-similarity.

$$SS_X = \int_{i \in D} \left| \frac{\partial}{\partial s} p_D(s, X) \right| di \tag{2}$$

To prevent deleting informative patches by mistake, we use dual threshold method. Only those patch whose self-similarity is smaller than $SelfSimilarity_{low}$ and local entropy is also smaller than $Entropy_{high}$ will be removed. The proper values of these thresholds are discussed in section 3.3.

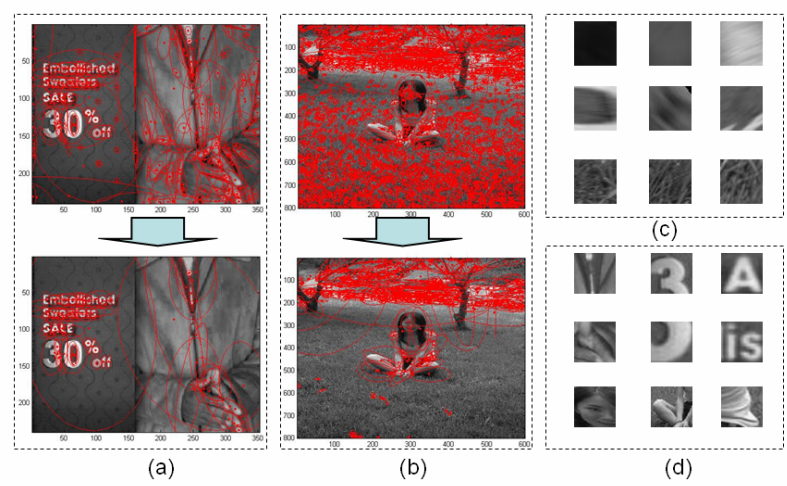

Figure 3. Samples of region filtration. (a) (b) Result comparison of region filtration.(c) Samples of removed background patches. (d) Samples of reserved patches.

3.2 Result on region filtrations

In this sub-section, the experimental result of patch filtration is discussed in detail. The image dataset used here are keyframes extracted from TRECVID 2005 news video retrieval database. Out of which 3000 images are selected. According to the objects they contain, these images are divided into 50 categories. The number of relevant images in each class ranges from about 20 to about 50 images,

while the rest are thought to be disturbances. All the subsequent experiments are based on MS region and SIFT descriptors.

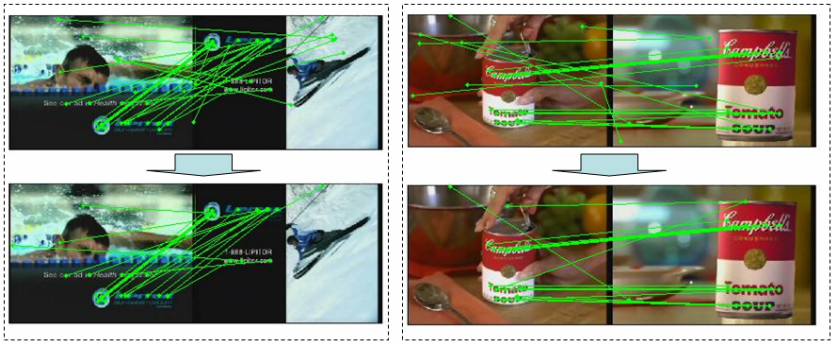

Figure 4. Examples for improved matching precision based on region filtration.

To measure the efficiency of patch filtration in section 3.2, we adopt exact point-to-point matching with SIFT in this sub-section. As shown in Figure 4, most of the false matching due to background patches are removed correctly.

There are 3 thresholds in our filtration process, among which $Entropy_{\text{low}}$ influences the performance mostly. So we test its influence in a sample image set including 200 images and about 28k patches are extracted, while we define $SelfSimilarity_{\text{low}}$=0.5 and $Entropy_{\text{high}}$=3.0. The influence on patches quantity and matching precision (the ratio of correct matching and all matching pairs found in each image) are shown separately in table 1 and Figure 5. We can see that when $Entropy_{\text{low}}$=2.0, the best balance can be achieved, while a lot of redundant patches can be removed correctly, and matching precision would be guaranteed at the same time. Accordingly, these filtration thresholds are adopted throughout our subsequent experiment.

Table 1. *Comparison of patches quantity*

$Entropy_{\text{low}}$	Delete amount	Delete ratio
1.0	589	2.1%
1.5	2407	7.3%
2.0	3786	13.5%
2.5	5076	18.1%

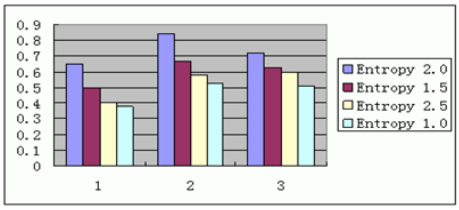

Figure 5. Comparison of matching precision

4 Spatial configuration model based re-ranking

Although region filtration can effectively remove a lot of background patches, the result of object retrieval still has some false matching due to the lack of spatial relation of these patches. In our system, we solve this problem with a novel method called spatial configuration model. The idea is implemented here by first retrieving query object using BoW method based on the selected regions, to obtain a small candidate image set, and then re-ranking them using the spatial model.

4.1 Spatial configuration model

The key issues of spatial re-ranking are spatial configuration definition and similarity measurement based on it. The following sub-sections describe our spatial configuration model in detail.

4.1.1 Spatial configuration definition

(J.Sivic, 2003; J.Phibin, 2007) defined the spatial configuration by the 15 nearest neighbours of each match using L2 distance (called L2-KNN method), and each patch which also matches with this area casts a vote for that matched patch. This definition doesn't consider the affine transformation of different images, and can't obtain the same spatial neighbouring area exactly. Here we define an affine covariant spatial configuration called Affine-KNN. According to the ellipse's parameter of central patch, those patches which lie in K (the experiential K in our system is 3) times the ellipse area are considered the spatial neighbours of this central patch. For example in figure 6(a), after affine transformation, L2-KNN mistakes the original neighbouring patch "b,c" for "d,e"(denoted with dashed circle), while our Affine-KNN method retains the original patch set exactly.

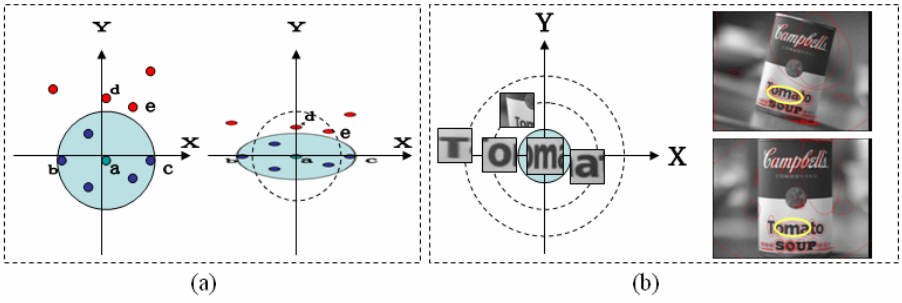

Figure 6. Spatial configuration definition. (a) Comparison of L2-KNN and Affine-KNN. (b)Example of normalized Affine-KNN spatial configuration in real image .

4.1.2 Spatial configuration similarity measurement

Based on the definition, configurations of a pair of matched patch i, j can be recorded as two patch sets: $S1 = \{c_{11}, c_{12}, ..., c_{1m}\}, S2 = \{c_{21}, c_{22}, ..., c_{2n}\}$,where c_{ij} is the visual word each neighbouring patch belonging to, and m,n denote the number of patches in each spatial configuration set. Note here the set also contains the information of which spatial level the visual word belongs to (level is between 1,2…K). The size of spatial configuration set often varies due to different image scale. For instance, image with small resolution has less detail and generates fewer patches. Consequently, the distance between two spatial configuration sets is measured using the Earth Mover's Distance (EMD) (Rubner *et al*, 1998). Considering that EMD matches perceptual similarity well and can operate on variable-length representations of the distributions, it is suitable for our spatial configuration similarity measure. Based on the spatial configuration definition and similarity measure, we define a pair of patches with spatial configuration similarity more than T_{emd} as "spatial matched patches", and re-rank the candidate image set use these spatial score.

4.2 Result on Spatial re-ranking

To evaluate the effectiveness of our spatial re-ranking method in object retrieval, 20 categories are selected as query images in the image dataset mentioned above which includes about 3000 images, and the number of relevant images in each class ranges from about 20 to 50.

Some examples for object retrieval result are shown as Figure. 7, where query objects are demarcated using yellow rectangle in the left query images. The top 5 images of retrieval result are shown with descending object's similarities. We calculate the time used to retrieve the top 20 relevant images to each query object. Using a 3.2G Pentium 4 PC with 1.5G memory, the average retrieval time for each query ranges from 0.11 second to 1.83 second depending on the number of visual words in the query object.

Figure 7. Examples for object retrieval result.

We compare our spatial configuration model based approach (called SCM) with J.Sivic's method (called L2-KNN) and No-spatial result (BoW without spatial information). The effectiveness of each approach is judged by a score as Q.F.Zheng's method, which is defined as follows:

$$Score(I_1,...,I_{20}) = \sum_{i=1}^{20} w_i X_i \tag{3}$$

$$w_i = \begin{cases} 2.0 & 1 \le i \le 5 \\ 1.5 & 6 \le i \le 10 \\ 1.0 & 11 \le i \le 15 \\ 0.5 & 16 \le i \le 20 \end{cases} \quad \text{and} \quad X_i = \begin{cases} 1, & I_i \text{ contains query object} \\ 0, & I_i \text{ contains no query object} \end{cases} \tag{4}$$

Where Ii is the top i-th retrieved image to a query object, and the weight wi is defined as (4). The average retrieval performances of 20 classed of the three approaches are plotted in Figure 8. As shown below, our SCM-based method generally outperforms L2-KNN based approach, and both of them are much better than visual word-based approach without spatial information. We attribute this to that spatial configuration model contains abundant spatial information between patches, and EMD-based similarity measurement is more suitable than simple voting method. However, compare the best class "News" with the worst class "man", we can find that our method is more adopted for dense patch images

such as “News” which including many distinctive and adjacent patch pairs, while in “man” there are few distinctive patches and they are often far from each other.

5 Conclusions and future work

We have presented an approach based on attention analysis and spatial re-ranking for object retrieval. In this system, a novel framework is proposed, and attention model is used to filter those background patches. Furthermore, The spatial relations of adjacent patches are described exactly and measured with EMD distance. Based on this, we use re-rank the BoW result with this information. Experimental result demonstrates that our method is efficient and outperforms the state-of-art object retrieval methods. We are currently investigating the extension of our proposed framework in the following aspects such as the amalgamation of more types of features and the use of relevant feedback.

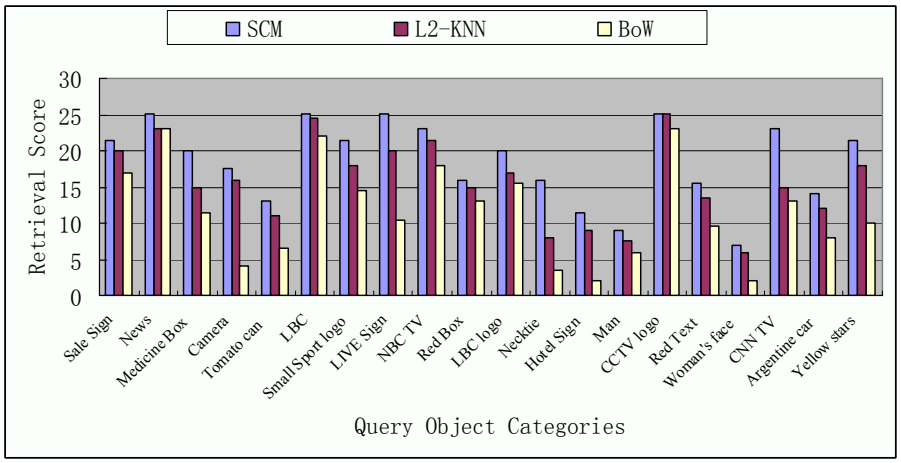

Figure 8. Average retrieval precision comparison.

6 Acknowledgements

This work was supported in part by the National High Technology and Research Development Program of China (863 Program, 2007AA01Z416) and the National Basic Research Program of China (973 Program, 2007CB311100), the National Nature Science Foundation of China (60773056), the Beijing New Star Project on Science & Technology (2007B071).

7 References

1. C. Carson, et al. Blobworld: A System for Region-based Image Indexing and Retrieval, In 3rd Int. Conf. on Visual Information Systems, 1999, Amsterdam, p. 509-516.
2. Itti L, Gold C, Koch C, Visual attention and target detection in cluttered natural scenes. Optical Engineering, 2001, Vol 40(9), p.1784-1793.
3. J. K. Tsotsos, S. M. Culhane, W.Y.K. Wai, et al, Modeling visual attention via selective tuning, Artificial Intelligence, 1995, p.507-545.
4. Jams Phibin, Ondrej Chum, Michael Isard, et al. Object retrieval with large vocabularies and fast spatial matching, IEEE Conference on Computer Vision and Pattern Recognition, CVPR2007.
5. J.Sivic, A.Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos, International Conference on Computer Vision, ICCV2003, p.1470-1477.
6. J.Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions, British Machine Vision Conference, BMVC2002, p384-393.
7. K.Mikolajczyk, T. Tuytelaars, et al. A comparison of affine region detectors, International Journal on Computer Vision, IJCV2006, p. 43-72
8. K.Mikolajczyk, C. chmid. A performance evaluation of local descriptors. IEEE Transaction on Pattern Analysis and Machine lntelligence, PAMI2005, p.615-1630.
9. Lowe, D. Distinctive image features from scale-invariant keypoints, International Journal on Computer Vision, IJCV2004, Vol 60(2), p.91–110.
10. Ma Y F, Zhang H J, Contrast-based image attention analysis by using fuzzy growing. Proceedings of the 11th ACM International Conference on Multimedia, MM2003. Berkeley, CA, USA: ACM, p.374 – 381.
11. Mikolajczyk, K. and Schmid, C. “An affine invariant interest point detector”, In Proceedings of the 7th European Conference on Computer Vision, ICCV2002, Copenhagen, Denmark.
12. Qing-Fang Zheng, et al. Effective and efficient object-based image retrieval using visual phrases, 14th ACM International Conference on Multimedia, MM2006, Santa Barbara, USA, p.77-80.
13. Rubner, Y., Tomasi, C., and Guibas, L., A Metric for Distributions with Applications to Image Databases. Proceedings of the IEEE International Conference on Computer Vision, ICCV1998.
14. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2006.
15. Timor Kadir, Michael Brady. Saliency, Scale and Image Description, International Journal of Computer Vision, IJCV2001. Vol45 (2), p.83-105.
16. Wang, J. Z., Li, J., and Wiederhold, G., SIMPLIcity: Semantics-sensitive Integrated Matching for Picture Libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI2001, vol. 23.